Evaluating the Trustworthiness of Wikipedia Articles through Quality and Credibility

Sai T. Moturu
Computer Science and Engineering
Arizona State University
Tempe, AZ, 85287
smoturu@asu.edu

Huan Liu
Computer Science and Engineering
Arizona State University
Tempe, AZ, 85287
huan.liu@asu.edu

ABSTRACT

Wikipedia has become a very popular destination for Web surfers seeking knowledge about a wide variety of subjects. While it contains many helpful articles with accurate information, it also consists of unreliable articles with inaccurate or incomplete information. A casual observer might not be able to differentiate between the good and the bad. In this work, we identify the necessity and challenges for trust assessment in Wikipedia, and propose a framework that can help address these challenges by identifying relevant features and providing empirical means to meet the requirements for such an evaluation. We select relevant variables and perform experiments to evaluate our approach. The results demonstrate promising performance that is better than comparable approaches and could possibly be replicated with other social media applications.

Categories and Subject Descriptors

H.3 [**Information Systems Applications**]: Miscellaneous

Keywords

Wikipedia, social media, trust, trustworthiness, quality

1. INTRODUCTION

The advent of Web 2.0 has changed the average web user from a consumer to a content creator. Social media applications like blogs, wikis and social networks are generating a large expanse of content, driven by user contributions. Wikipedia is one of the more famous examples of social media. The social web has its share of useful knowledge in combination with untrustworthy content. Search engines are the gateways to such content. However, search relevance does not indicate whether the content is trustworthy. With the advent of social media content contributed by unknown users, it is necessary to assess the trustworthiness of every piece of content as trust cannot be placed on a web portal as was the case earlier. The presence of a trust assessment, in addition to search relevance, can change the way people perceive and utilize information from social media. A number of works in recent years have quantitatively assessed quality and trust in Wikipedia articles. In this work, we differentiate between quality and trust. Further, we believe content, revision history and author information are all helpful in our quest for trust assessment, unlike many previous works that use only one aspect of the data.

WikiSym '09 October 25–27, 2009, Orlando, Florida, U.S.A.

2. TRUST

Trust is an important sociological concept that has been studied in depth for a number of years. For the purpose of this paper, we rely on the terms *trust* and *trustworthiness* to focus on content reliability. Trust is a concept involving in a transaction between two entities, the trustor and the trustee. Trust can be defined as the perception of the trustor about the degree to which the trustee would satisfy an expectation about a transaction constituting risk. Trustworthiness can be defined from the perspective of both these entities. In this paper, we will only consider the perspective of the trustor, which defines this property to be the amount of trust associated with the trustee [1].

3. TRUST ASSESSMENT

We select 230 health-related articles from Wikipedia including articles tagged as Featured, Good, Cleanup and Stub. These articles are manually classified by Wikipedia users based on predefined requirements. Table 1 describes the data distribution. Featured articles are of the highest quality, closely followed by good articles. Cleanup articles are specifically marked for improvement while stubs have minimal content. Articles that do not fall under any of these categories are considered Standard articles.

Our approach to trust evaluation is divided into three major tasks. The first task is the identification of relevant features capable of assessing the reliability of content. Next is the creation of feature-driven trust evaluation models that are independent of the application. The final task is the performance evaluation of these models.

3.1 Feature Identification

We first describe two categories of information from which relevant features can be derived and follow it up with a description of the selected features.

Quality. Quality represents an inherent feature or essential character. Predictors derived from content can be used to define quality. Quality is sometimes used interchangeably with trust but these issues are distinct [4].

Table 1: Wikipedia Data

Category	Featured	Good	Standard	Cleanup	Stub
Articles	25	25	105	25	50

Table 2: Experimental Results

Model	NDCG (top 50)	NDCG (all)
Dispersion	0.8343	0.9398

Features that can help us ascertain information provenance fall into the *source and citation quality* category. The first feature in this category is a weighted reference score per paragraph where a higher weight is assigned to peer-review publications. The second feature is the proportion of paragraphs with citations. External characteristics are another source for features useful in the assessment of *content quality*. Article size is a suitable feature for this category.

Credibility. Credibility is the quality of inspiring belief. Factual accuracy is a suitable property of reliable content. Metadata associated with the content can include information on editing patterns, development history and user behavior, among other things. Predictors derived from such metadata associated directly or indirectly with the content measure the credibility of an article.

As social media is all about user-driven content, it is important to assess *author credibility*. Information about user activity is available across social media applications. Three features are selected for this category: proportion of unregistered editors with a single edit, weighted unregistered editor contribution score (higher weights for those with more contributions), proportion of registered editors with mean edit frequency under six hours. The development of an article, the response to it and audience participation can all be used to assess *formative credibility*. Features in this category include: revision count, proportion of reverted edits, proportion of reverted edits considered vandalism, mean time between successive edits and mean edit length.

3.2 Trust Evaluation Model

We propose the dispersion degree score (DDS) model to assess and score the trustworthiness of articles. In this model, the dispersion of the feature values from their mean is utilized to derive their relative importance. The underlying assumption is that the farther a feature value is from its mean, the greater its effect on trust.

$$S_{ij} = \begin{array}{c} 0 \; if \; f_{ij} < m_i - d_i \\ x+1 \; if \; m_i - d_i + cxd_i < f_{ij} < m_i - d_i + c(x+1)d_i \\ 11 \; if \; f_{ij} < m_i + d_i \end{array} \quad (1)$$

$$T_D(i) = \sum_{j=1}^{n} S_{ij} \quad (2)$$

A score, S_{ij} is assigned to each feature f_{ij} based on the dispersion of its value from the mean, m_i as measured by the standard deviation d_i (eq. 1). Each feature can fall in one of twelve trust classes with scores from 0 to 11. The constant c is used to define the class interval and a value of 0.2 is used here. The sum of these scores provides the final trust score (eq. 2) where a larger score indicates a more trustworthy article.

3.3 Evaluation Methodology

The Normalized Discounted Cumulative Gain (NDCG) evaluation metric [3] is used to evaluate the performance of our models. The trust scores output from each of our models can used to rank articles. The classification of the Wikipedia articles in our data can be ordered by reliability: Featured > Good > Standard > Cleanup > Stub. For our evaluations, we use a tie-oblivious version of NDCG [5] that takes tied scores into account.

4. EXPERIMENTAL RESULTS

Table 2 depicts the experimental results. To provide a comparison, we also calculate baseline performances for the worst and average cases. The worst case occurs when the articles are ranked in reverse. For the top 50 articles, this value is 0.001 and for all articles, it is 0.45. For the average case, when all the articles are ranked the same, this value is 0.21 for the top 50 and 0.59 for all articles.

It is easily apparent that our model is much better than the average case when it is assumed that all articles are equally reliable. These results are impressive. The quality models proposed by Hu et al. [2] result in lower NDCG ($<$ 0.9 for all articles, k=242 and $<$ 0.8 for k=45) for a different data set with a similar number of Wikipedia articles. Despite differences, this comparison indicates that our models perform at least as effectively, if not more.

5. CONCLUSIONS AND FUTURE WORK

In this work, we define the notion of trust in terms of quality and credibility and use it to formulate a broad framework to assess the trustworthiness of Wikipedia content. Promising experimental results render our approach sound. Our future work includes an improvement of our current model and the extension of this work to other social media.

6. ACKNOWLEDGMENTS

This work is sponsored, in part, by grants from ONR (N000140810477) and AFOSR (FA95500810132).

7. REFERENCES

[1] B. Bailey, L. Gurak, and J. Konstan. Trust in cyberspace. *Human factors and Web development*, pages 311–21, 2003.

[2] M. Hu, E. Lim, A. Sun, H. Lauw, and B. Vuong. Measuring article quality in wikipedia: models and evaluation. In *Proc. of the 16th ACM conf. on information and knowledge management*, pages 243–252. ACM, NY, USA, 2007.

[3] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

[4] K. Lampe, P. Doupi, and J. van den Hofen. Internet health resources: from quality to trust. *Methods of information in medicine*, 42(2):134–142, 2003.

[5] F. McSherry and M. Najork. Computing information retrieval performance measures efficiently in the presence of tied scores. *Lecture Notes in Computer Science*, 4956:414, 2008.